

本周周报

解聪

2013.12.30-2013.01.05

本周工作

1. 完成毕业论文的初稿

2. 和尚平平一起总结了类别型数据探索到目前的工作进展，写在毕业论文中。

2000 KDD 的一篇文章和我们的工作相似：Visualizing Association Rules with Interactive Mosaic Plots

文章指出关联规则的两个问题：规则太多，人无法大量处理；规则难以理解。（2000 年提出的，可能现在有新问题）。

这篇文章利用可视化解决问题二。设 x, y 是一组变量的具体取值，这篇文章重点对 $x \rightarrow y$ 这样的规则进行可视化并结合上下文进行解释。受可视化方法限制，该文章的 y 只能取某一个具体属性值。

定义 x 是一组条件值， y 是目标值。我感觉我们方法能做的事情比这篇文章要多。

在我们的工作中，我们不仅可以对 $x \rightarrow y$ （如商品 \rightarrow 人群）这样的规则进行可视化和联系上下文解释和过滤；还可以对不同条件值 x_1 和 x_2 （如 qq 专区和数据的关系）进行解释和修改；以及不同目标值 y_1, y_2 之间的关系（如年轻男性和中年女性）的关系进行解释和修改。

不过我们还是需要在这篇文章的 100 多篇引用文献中寻找相似的工作，看我们的想法是不是已经被别人实现了。

下周工作：

研究经济数据的一些数据模型，看看能否用在本文系统上。

深入探索关联规则挖掘的可视分析的相关工作。

（下面附上论文中类别型数据探索的相关章节：）

4 用户交易类别型数据的可视化

4.1 本章可视分析方法概述

类别数据，包括频率数据和离散数据，几乎存在与全世界所有表格中。分析师通常处理这些数据时会使用Loglinear模型和Logit模型等方法^[58]以用来解决参数估计的问题。而过去一段时间，针对类别型数据的方法迅猛发展。相关可视化的工作及各种方法针对的不同数据应用场景的分析在第1.2.3节中已有介绍。

本节针对用户交易数据中的多维类别型属性进行分析。类别数据可以理解为是用户自身的不同标签属性，比如性别，年龄等。类别数据同时还包括用户的购买属性，比如购买商品的类目等。通过对这些类别型数据的分析，我们可以发现不同属性之间的关联以及相同属性不同取值的关联。本文将问题概括为对多维离散随机变量的分析。本文方法将不同的类别属性看成离散随机变量，而每个用户个体的数据则是离散随机变量的实例。

4.2 任务定义

本文重点对交易数据中的类别属性进行分析。我们使用了包含约1万个用户的历史交易数据，每个用户的属性可以分为两类：一类是基本属性，包括用户的性别，出生年份，星座等等。另一类是行为属性：包括用户的购买商品的类目，买家等级，买家年限等等。分析师一般关心的问题包括：

- 具有不同基本属性的用户群其行为属性的特征。比如不同年龄段、性别用户经常购买哪些类目的商品。
- 具有不同行为属性的用户群其基本属性的特征。比如某一特定商品是经常被那个年龄段，那种性别用户购买。
- 查询具有相似属性特征的用户群。比如查询哪些用户群具有与20岁左右的男性用户相似的购物倾向。

- 修改不同用户群之间的关联。有时候自动算法会产生一些误差，这时候需要分析师交互式地调整数学模型。
- 提取用户群或商品之间的关联。

由以上的交易数据案例的一些分析任务，本文对多维离散随机变量的可视分析的问题进行概括。并重点分析如下几方面的问题：

- 对不同离散随机变量组合的联合概率分布，条件概率分布的可视化。
- 对不同条件下相似的条件概率分布的查找。
- 对不同条件下离散随机变量的概率分布的相似性的修改与重定义。
- 对离散随机变量的概率分布的特征的提取。特征包括：不同分布全局的相似，互补，以及更为复杂的局部的相似性。

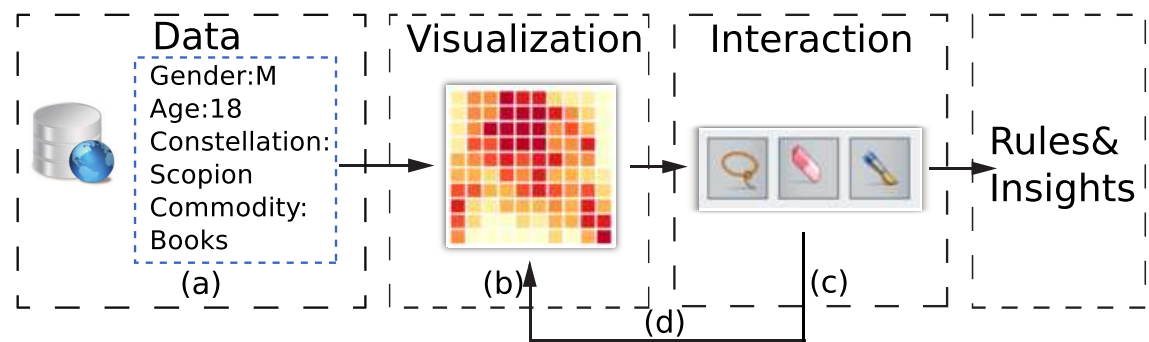


图 4.1 多维离散型随机变量的可视化流程。

本节对多维随机离散变量数据可视分析方法的系统流程如图4.1所示。

4.3 离散随机变量的矩阵生成

相关工作已在本文的1.2.3中有介绍。本文方法改进了Mosaic图的不足，并结合了small-multiple和矩阵图以表达不同类别的数据。Mosaic图将矩形划分为多个Block以展示类别属性的组合。它对每个划分后的Block迭代划分导致了多次划分后的Block布局不一致，从而没有办法进行直接比较。另外由于划分的区域大小编码了数据属性，所以不均匀的划分导致了有一些划分后的Block太小不便于观察。

Algorithm 4 划分算法

Input: 离散随机变量集合: $A = \{A_1, A_2, \dots, A_k\}$, 其中 A_i 的取值个数为 n_{A_i} 。矩阵中被原始未被划分为的方块的序列 $BLOCK = \{block_1\}$ 。Block矩阵行数 $N_{BlockRow} = 1$, 列数 $N_{BlockCol} = 1$ 。

```

1: for Each  $A_i$  in the set  $A$  do
2:   指定横向划分或纵向划分的方式;
3:   for Each  $block_{x,y}$  in the set  $BLOCK$  do
4:     在横向或纵向上将 $block_{x,y}$ 分裂为 $n_{A_i}$ 个block;
5:     if 对block进行横向划分 then
6:       计算新的行数  $N_{BlockRow} = N_{BlockRow} \times n_{A_i}$ ;
7:     else
8:       计算新的列数  $N_{BlockCol} = N_{BlockCol} \times n_{A_i}$ ;
9:     end if
10:    对新的矩阵中的block重新编号;
11:   end for
12: end for

```

我们仍然按照Mosaic图的划分方式, 将矩形划分为Block矩阵。设数据中的全部 n 个离散随机变量为 $V = \{V_1, V_2, \dots, V_N\}$, 其中每个离散变量 V_i 的, 其取值个数为 n_{V_i} 。我们使用一组 k_1 个随机变量序列 $A = \{A_1, A_2, \dots, A_{k_1}\}$ 先对矩形进行划分。划分算法如4所示。

由划分算法4可得划分后Block矩阵行列数的计算公式4.1。

$$N_{BlockRow} = \prod n_{Ar_i} \quad (4.1)$$

$$N_{BlockCol} = \prod n_{Ac_i} \quad (4.2)$$

其中 Ar_i 和 Ac_i 表示水平划分和纵向划分的变量, n_{Ar_i} 和 n_{Ac_i} 表示这些离散变量的取值个数。

对于每一个Block, 我们仍然可以使用一组 k_2 个随机变量序列 $B = \{B_1, B_2, \dots, B_{k_2}\}$ 对每个Block进行划分。得到更细的单元Cell组成的矩阵。我们采用和Block矩阵的生成相同的划分方式生成Cell矩阵, 只需在算法4中将划分对象换为Block, 划分结果由Block矩阵换为Cell矩阵即可。设Cell矩阵的行数为 $N_{CellRow}$ 和列数 $N_{CellCol}$ 。则由算法可以推出Cell矩阵行列数的计算公式4.3

$$N_{CellRow} = \prod n_{Br_i} \quad (4.3)$$

$$N_{CellCol} = \prod n_{Bc_i} \quad (4.4)$$

其中 Br_i 和 Bc_i 表示水平划分和纵向划分的目标维度， n_{Br_i} 和 n_{Bc_i} 表示这些离散维度的取值个数。

在下文中，我们将离散随机变量集合 A 称作条件变量，其取值 $a_1, a_2, a_3, \dots, a_{k1}$ 称作条件值；将离散随机变量集合 B 称作目标变量，其取值 $b_1, b_2, b_3, \dots, b_{k2}$ 称作目标值。

4.4 矩阵可视化方案

为了弥补划分后Block和Cell大小一致，而无法编码划分后的Cell数据量的信息，我们使用Cell的颜色编码数据信息。

每个Block对应了 A 的一组特定取值 $s_1 = \{a_1, a_2, a_3, \dots, a_{k1}\}$ 。则每个Block中实例的个数在总数据中的比例可以近似认为是联合分布的概率分布 $P(a_1, a_2, a_3, \dots, a_{k1})$ 。特别的，Block矩阵中某一行或某一列所对应的实例个数在总体数据中的比例就是联合分布的边缘概率分布。

对Cell的编码可以有以下三种方式：

4.4.1 条件值和目标值的联合概率可视化

而每个Cell也对应了在一组离散随机变量 A 取值 s_1 的情况下， B 的一组特定取值 $s_2 = \{b_1, b_2, b_3, \dots, b_{k2}\}$ 。则每个Cell的实例个数表示了联合分布 $P\{(A = s_1), (B = s_2)\}$ 。可以将该联合概率映射到颜色编码上，反映目标值和条件值的支持度。这种方案同时可以认为是对条件变量集合 A 和目标变量 B 构成的数据的列联表的可视化。

4.4.2 条件值和目标值的条件概率可视化

而每个Cell也对应了在一组离散随机变量 A 取值 S_1 的情况下， B 的一组特定取值 $s = \{b_1, b_2, b_3, \dots, b_{k2}\}$ 。则每个Block中的Cell表示了条件概率分布 $P\{(B = s_2)|(A = s_1)\}$ 。将条件概率映射到颜色编码上，以反映目标值和条件值的关系，即目标值对条件值的置信度。

4.4.3 关联规则的可视化

为了发现数据的关联规则，即 $s_1 \rightarrow s_2$ 。本文方法计算了规则的置信度 $P(s_2|s_1)$ ，和支持度 $P(s_1, s_2)$ 。我们由分析师设定置信度和支持度的阈值。对于置信度和支持度均高于阈值的 s_2, s_1 ，本文方法对相应的Cell进行高亮显示，对目标值和条件值之间的关联进行可视化。通过交互，改方法可以帮助分析师进一步理解规则的内在含义。

4.5 交易可视化案例

本节结合具体的交易案例解释以上矩阵划分算法。我们在交易数据中以用户为单位，探索用户类别型属性之间的关联。我们先选取条件变量集合为 $A = \{A_1\}$ ，其中 A_1 为交易类目。目标变量集合为： $B = \{B_1, B_2, B_3\}$ ，其中 B_1, B_2, B_3 分别为性别，年龄和星座。我们的目的在于分析不同交易类目与不同用户基本属性的关联，以及不同类目在不同用户群上分布的相似性。我们选取了购买商品进行初始划分。其次我们对于每个划分后的区域，使用年龄和星座依次横向划分，使用年龄纵向划分。

我们使用条件属性 A 对原始数据进行划分，得到一个 $1 \times n_{A_1}$ 列的Block的矩阵。划分的效果如图4.2所示。

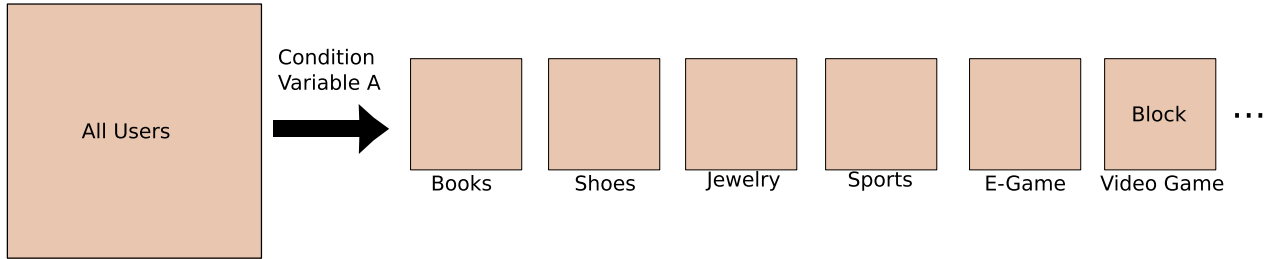


图 4.2 利用条件变量划分矩形，得到Block的矩阵。

对于每一个Block，我们使用目标属性集合 B 对其进行划分，得到一个Cell矩阵。划分的效果如图4.3所示。

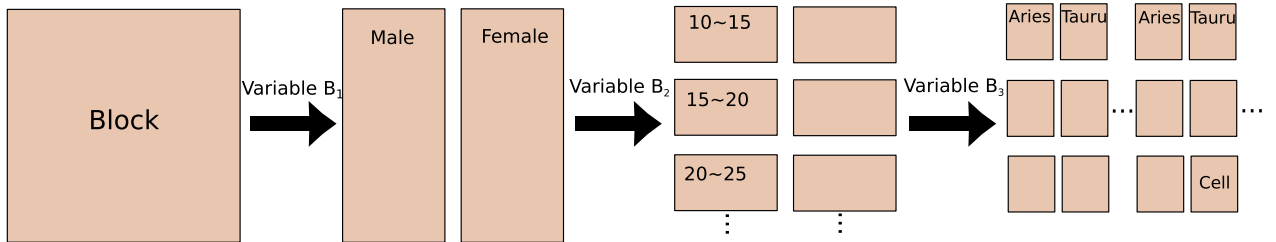


图 4.3 对每个Block的划分示意图，得到Cell的矩阵。

我们使用颜色编码每个Cell中的条件概率值。概率越大，颜色越红，得到的结果如图4.4所示。由于 $1 \times n_{A_1}$ 的Block的矩阵太长，我们这里在图中将Block的矩阵折叠显示。

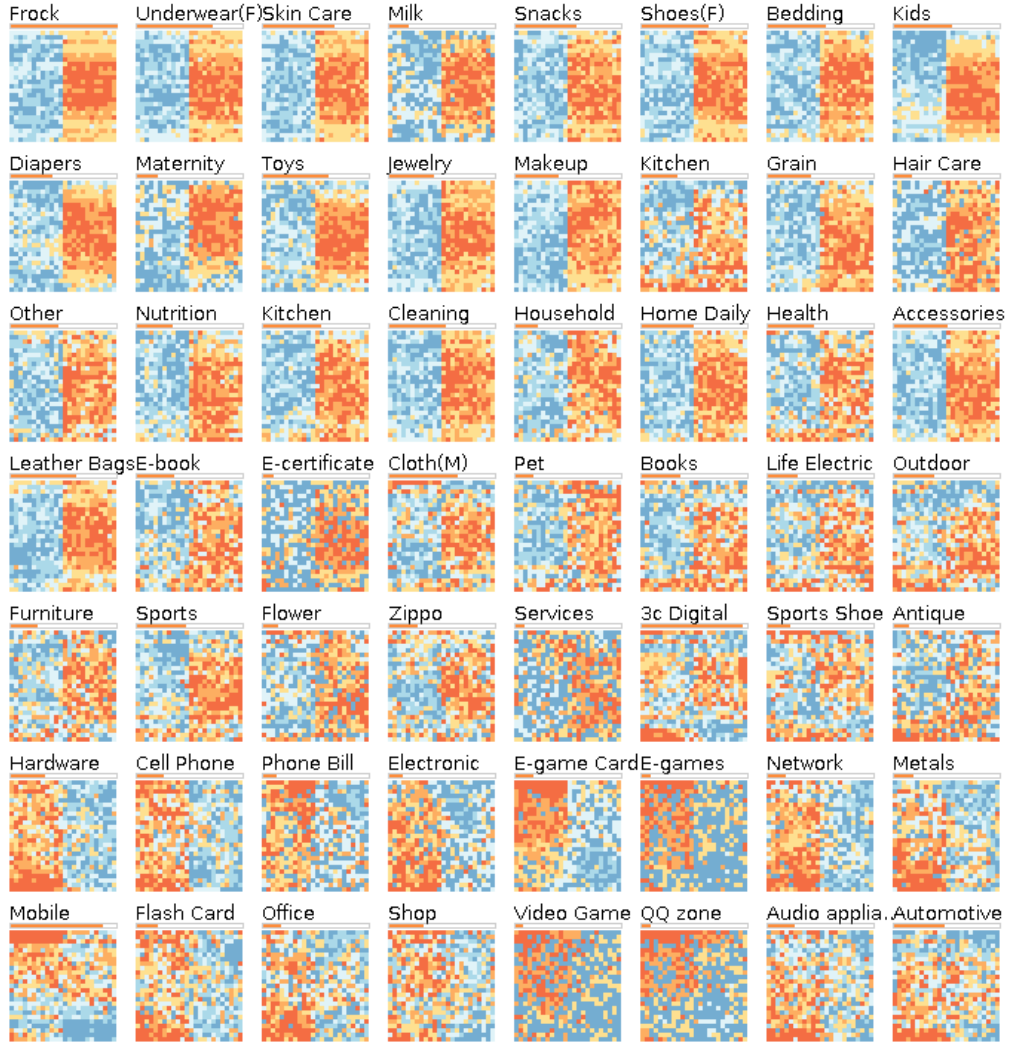


图 4.4 本节方法对用户交易数据的可视化，其中每一个方块代表变量 A_1 交易类目的一种取值，方块内部横向按照变量 B_1 性别和变量 B_3 星座划分，纵向按照变量 B_2 年龄划分。

图中概率密度的从小到大映射到蓝色到红色。单个方块中的概率密度函数反映了购买某一商品的用户人群的分布情况。可以看出图中有些交易模式非常明显，比如 $A_1 = \text{鞋子“Shoe”}$ 和珠宝“Jewelery”等交易类目对应的Block基本上都是性别属性 $B_1 = \text{女性}$ 的Cell购买密度比较大。而 $A_1 = \text{电子游戏“E-Game”}$ ，视频游戏“Video Game”等交易类目的对应的Block中， $B_1 = \text{男性}$ 群体的Cell购买密度较大。特别的，我们注意到男性群体中， $B_2 = \text{年轻男性}$ 的购买概率要远大于 $B_2 = \text{老年男性}$ 的用户。可以猜测用户群体中青少年男性特别喜欢玩网游，而他们在 $A_1 = \text{运动“Sports”}$ 等类目中对应的分布概率密度很小。因此，我们可以猜测这一用户群体很少外出运动，而经常宅在住所玩网络游戏。

4.6 相似性的定义

从图4.4中可以发现,不同类目的相似性可以由图像模式的不同视觉上直接发现。因此,我们这里将目标随机变量 B 在两组条件属性值 a_1, a_2 下的分布的相似性定义为公式。

$$Sim(a_1, a_2) = \sqrt{\sum_{i=1}^n w_i \times (P\{(B = b_i)|a_1\} - P\{(B = b_i)|a_2\})^2} \quad (4.5)$$

需要说明的是,这里的 w_i 表示 B 取 b_i 时对应的权重。将 w_i 排布在一个 $n \times n$ 的零矩阵 O 的对角线上,得到其权重矩阵 W 。在现实分析中,我们发现相似性的计算可能收到领域知识的影响。比如在图4.4中,如果分析师只关注 $B_1 = \text{女性}$ 的用户的购买行为,可以将 $B_1 = \text{男性}$ 用户的权重置为0,进而重新分析不同类目属性 A_1 取值的相似关系,即类目的关联分析。

4.7 交互式探索工具

我们重点关注被划分到各个Block和Cell中的数据分布的特征。进而发现目标属性 B ,条件属性 A 之间的关联,以及目标属性 B ,条件属性 A 各自内部的相似性与关联。

我们将交互总结归纳为5类,其交互方式和对应于数学模型的操作如下:

4.7.1 Block和Cell的选取与标记

针对不同的选取对象,选取和标记的操作可以分为如下几类:

- 对一个Block的选取与标记: 用户通过鼠标单击 $block_{x,y}$ 。设 $block_{x,y}$ 对应的条件属性的值集合为 $s = \{a_1, a_2, a_3, \dots, a_{k1}\}$, 则令条件变量 $A = \{A_1, A_2, \dots, A_k\}$ 的对应取值为 s 。
- 对矩阵图中一组Block的圈选: 用户通过鼠标拖动划选一系列的Block集合。设其选择的Block所对应的条件值集合为 $S = \{s_1, s_2, s_3, \dots, s_n\}$, 则令条件变量 A 的取值组合为 S 中的任意个元素 s_i , 即 $s_1 \vee s_2 \vee s_3 \vee \dots \vee s_n$ 。
- 对Block整行(列)的选取: 用户通过点击选取Block矩阵中的第 y 行 $Block_{i,y}$, ($0 \leq i \leq N_{BlockCol}$)。该行对应的条件值集合为 $S = \{s_1, s_2, s_3, \dots, s_m\}$, m 为纵向划分矩阵的条件变量的个数。则令条件变量 A 的取值组合为 S 中的任意个元素 s_i , 即 $s_1 \vee s_2 \vee s_3 \vee \dots \vee s_m$ 。列的选取也以此类推。

- 对Cell的圈选：对矩阵图中一组Cell的圈选：用户通过鼠标拖动划选一系列的Cell的集合。设其所处的Block的条件值集合为 $s' = \{a_1, a_2, a_3, \dots, a_{k1}\}$ 。某个 $Cell_{x,y}$ 所对应的目标值集合为 $s = \{b_1, b_2, b_3, \dots, b_{k2}\}$ ，其选择的Cell集合所对应的目标值集合为 $S = \{s_1, s_2, s_3, \dots, s_n\}$ 。则令条件变量 A 取 s' ，目标变量 B 的取值组合为 S 中的任意个元素 s_i ，即 $s_1 \vee s_2 \vee s_3 \vee \dots \vee s_n$ 。此时当圈选个数 $n = 1$ 时，该操作变为对Cell的单选。

我们设计了一套对矩阵图进行选取的工具，以方便用户的圈选，框选和单个点击选取的交互，如图4.5所示。

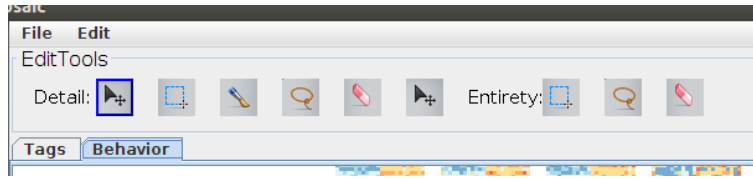


图 4.5 系统对矩阵图的探索工具。包括划选，点击，套索等。

4.7.2 行列的过滤删除

我们可以通过对Block整行（列）的选取选定的行和列。由于某些属性的组合可能在实际数据中无有意义，比可能存在如 $P(A_1 = a_1, A_2 = a_2) = 0$ 的情况。此时划分后的包含有 a_1, a_2 的所有Block所对应的实例数目为0。这种情况下可以将该行或列删除，以减少视觉冗余。

同时，我们可以认为是分析师想要观察所重点感兴趣的数据，所以会对Block进行整行（列）的选取。因此可以将无关的数据先暂时过滤掉，比如对应与 $A_1 \neq a_1, A_2 \neq a_2$ 的Block集合，以缩小数据的探索范围。被过滤的数据可以被重新添加到数据集中。

4.7.3 行列的重排列

矩阵中，不同行列的Block中数据分布情况不同，我们可以将行列进行重新排序使得相似的行列布局上靠近，以方便分析师的比较。设对矩阵进行纵向进行划分的条件维度集合为 A_1 ，其中某两行的Block的 A_1 条件属性的取值的集合为 s_1, s_2 。依据公式4.5。设对矩阵进行纵向进行划分的条件维度集合为 A_2 ，我们计算条件随机变量集合 A_2 在两组条件属性值 s_1, s_2 下的分布的相似性 $Sim(s_1, s_2)$ 。列的相似性计算与之类似。

本文将相似的行列自动重排到相近的位置。同时，对于行列的自动重排结果。用户可以对Block整行（列）选取并且拖动到相对应的位置，调整布局的结果。

4.7.4 Block相似性的查询

按照公式4.5对相似性的定义，本文支持在Block矩阵中寻找与指定 $Block_{x,y}$ 相似的Block的集合。对于选取的单个 $Block_{x,y}$ ，本文支持两种相似性的查询：

- 考虑Block中全部Cell的相似性计算。设 $Block_{x,y}$ 其条件属性的值为 $s = \{a_1, a_2, a_3, \dots, a_{k1}\}$ 。矩阵中任意一个 $Block_{x_i,y_i}$ 的条件属性值为 s_i ，则本文方法计算目标随机变量集合 B 在两组条件属性值 s_i, s 下的分布的相似性 $Sim(s_i, s)$ 。在计算过程中，权重矩阵 W 为单位矩阵 E ，每一个权重值对应于一组目标变量 B 的取值。我们选取相似性高于一定用户指定阈值的Block，提供给用户做进一步分析。
- 考虑Block中部分Cell的相似性计算。用户可以使用Cell矩阵的圈选操作（图4.6）选取一部分的Cell，选取的目标变量集合 B 对应的取值集合为 s' 。对于选中的Cell，将其在权重矩阵 W 对应的权重量为1，其他的置为0。设 $Block_{x,y}$ 其条件属性的值为 s 。矩阵中任意一个 $Block_{x_i,y_i}$ 的条件属性值为 s_i ，则本文方法计算目标变量集合 $B = s'$ 时，其在两组条件属性值 s_i, s 的分布的相似性 $Sim(s_i, s)$ 。此时按照修改后的权重值结果计算相似性。

4.7.5 Block相似性的修改

由于通过对Cell的简单划选，并计算局部的相似性的方法可能仍然无法满足用户的需求，因此我们提供进一步对权重矩阵进行修改的方法。这里我们采用距离尺度学习^[59]的方法，自动训练出两组随机变量取值的相似性。我们先通过Block的划选标记，将所有Block分类为 m 类 C_1, C_2, \dots, C_m 。对于任意的 $Block_{x_i,y_i}$ ，其的条件值为 a_i ，我们求解 W 。

$$\min_W \sum_{Block_{x_i,y_i}, Block_{x_j,y_j} \in C_k} Sim(a_i, a_j)^2 \quad (4.6)$$

$$s.t. \sum Sim(a_i, a_j)^2 \geq 1 \quad (4.7)$$

$$W \succeq 0 \quad (4.8)$$

由于此时 W 是对角矩阵，设

$$G(W) = \sum_{Block_{x_i,y_i}, Block_{x_j,y_j} \in C} Sim(a_i, a_j)^2 - \log(\sum Sim(a_i, a_j)^2) \quad (4.9)$$

可以通过Newton-Raphson方法直接求解 G 的最小值($W \succeq 0$)，来求解最优的权重矩阵 W 。

4.8 交互分析案例

我们注意到 $A_1 = \text{QQ专区“QQ Zone”}$ 的购买用户有比较特殊的模式。其中年轻的用戶购买的比较多，男女区别不是很明显，但是随着年龄的增加，女性对应的概率密度减小，而男性用户不变。当年齡继续增加的时候，可以发现男性用户的密度相对于年轻男性业主间减少。而书籍“Books”类目与之正好是完全相反的模式。随着年龄增大，对应用户的密度逐渐增大。

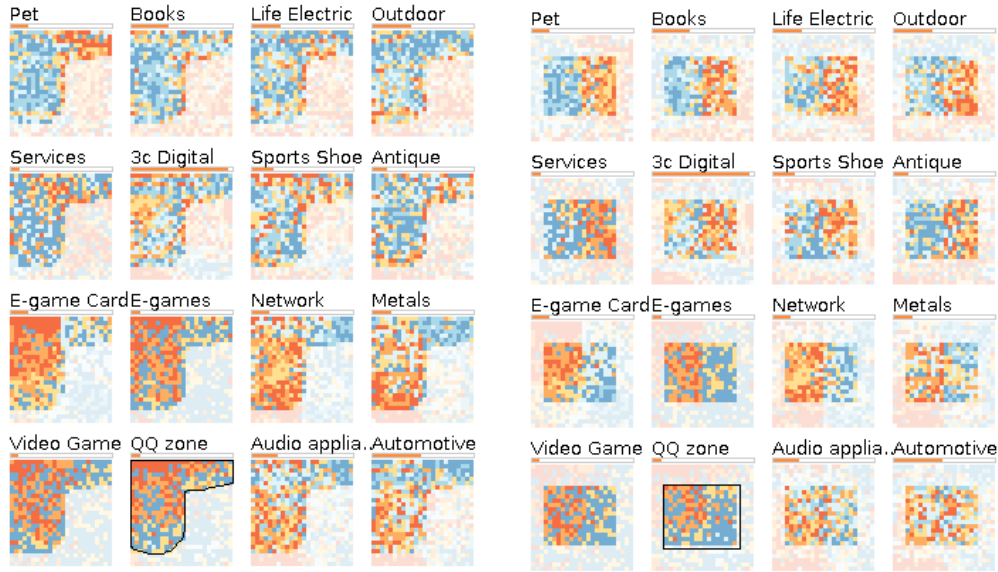


图 4.6 对交易数据进行套索操作和框选操作的可视化效果。

我们使用本章定义的交互工具帮助我们进行分析。我们可以使用套索工具（图4.6 左）选择出Block所呈现出来的不规则图像形状。我们选取了 $A_1 = \text{QQ专区“QQ Zone”}$ 类目的密度分布比较大的用户群。可以观察到该类目与 $A_1 = \text{书籍类目“Books”}$ 现互补趋势。我们可以筛选掉未被选中的用户群对应的Cell，重新计算不同类目的在这部分用户中的相似性。设选取的目标属性 B 对应的取值集合为 s' ，则计算在目标随机变量集合 $B = s'$ 时，在两组条件属性值“QQ Zone”和其他 $A_1 = a_1$ 下的分布的相似性 $\text{Sim}(\text{“QQ Zone”}, a_1)$ 。如果我们只关注 $B_2 = \text{中年用户的行为模式}$ ，我们可以使用框选交互（图4.6 右）选取该段人群，重新计算该段人群中不同商品类目在人群分布中的相似性。计算结果发现 $A_1 = \text{“QQ Zone”}$ 与古董“Antique”关联度最低，与“E-game Card”和“Video Game”关联度最高。